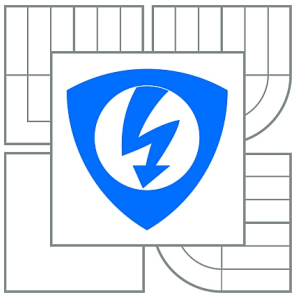
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF BIOMEDICAL ENGINEERING

# KLASIFIKACE ORGANISMŮ NA ZÁKLADĚ NUKLEOTIDOVÝCH ČETNOSTÍ

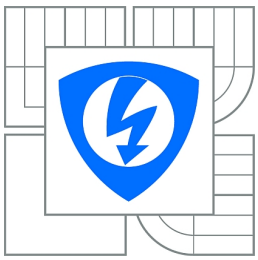
CLASSIFICATION OF ORGANISMS USING NUCLEOTIDES FREQUENCIES

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — NICOL ZBOŘILOVÁ
AUTHOR

VEDOUCÍ PRÁCE — Ing. HELENA ŠKUTKOVÁ
SUPERVISOR

BRNO 2014



**VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ**

**Fakulta elektrotechniky a komunikačních technologií**

**Ústav biomedicínského inženýrství**

# Bakalářská práce

bakalářský studijní obor
**Biomedicínská technika a bioinformatika**

***Studentka:*** Nicol Zbořilová ***ID:*** 147467
***Ročník:*** 3 ***Akademický rok:*** 2013/2014

**NÁZEV TÉMATU:**

## Klasifikace organismů na základě nukleotidových četností

**POKYNY PRO VYPRACOVÁNÍ:**

1) Seznamte se s problematikou vyhodnocení příbuznosti organismů na základě podobnosti DNA sekvencí. 2) Vypracujte literární rešerši metod vyhodnocujících příbuznost organismů na základě charakteristických nukleotidové četností. 3) Navrhněte a realizujte v programovém prostředí Matlab algoritmus pro klasifikaci organismů na základě specifické četnosti dinukleotidů a nukleotidových tripletů. 4) Vytvořte program s grafickým uživatelským rozhraním pro klasifikaci organismů formou fylogenetického stromu na základě alespoň tří metod využívajících nukleotidové četnosti. 5) Program doplňte o standardní vyhodnocení fylogenetického stromu z proporcionálních vzdáleností. 6) Program otestujte na vhodně zvolených sekvencích z veřejných databází. Proveďte srovnání všech realizovaných metod a výsledky diskutujte.

**DOPORUČENÁ LITERATURA:**

[1] QI, X., E. FULLER, Q. WU a C. Q. ZHANG. Numerical characterization of DNA sequence based on dinucleotides. ScientificWorldJournal, 2012, 2012, 104269.
[2] RANDIC, M., X. GUO a S. C. BASAK. On the characterization of DNA primary sequences by triplet of nucleic acid bases. J Chem Inf Comput Sci, May-Jun 2001, 41(3), 619-626.

***Termín zadání:*** 10.2.2014 ***Termín odevzdání:*** 30.5.2014

***Vedoucí práce:*** Ing. Helena Škutková
***Konzultanti bakalářské práce:***

**prof. Ing. Ivo Provazník, Ph.D.**
*Předseda oborové rady*

**UPOZORNĚNÍ:**

# ABSTRAKT

Tato práce se snaží představit různé přístupy analýzy genomických dat a klasifikace organismů. Chce porovnat účinnost klasických metod, založených na nutnosti vzájemného zarovnání sekvencí, které jsou tímto výpočetně náročnější s moderními přístupy, využívajícími pouze četnosti jednotlivých nukleotidů či jejich skupin v biologických sekvencích.



# KLÍČOVÁ SLOVA

Fylogenetika

Genomika

Sekvenace DNA

Zarovnání sekvencí

Klasifikace



# ABSTRACT

This thesis tries to present different approaches of analysis of genomic data and classification of organisms. This thesis also wants to compare the effectiveness of traditional methods based on the necessity of aligning sequences that are computationally demanding and modern approaches utilizing only the frequencies of individual nucleotides or groups of them in biological sequences.



# KEYWORDS

Phylogenetics

Genomics

DNA sequencing

Alignment of sequences

Classification

Zbořilová, Nicol *Klasifikace organismů na základě nukleotidových četností*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií. Ústav biomedicíny, 2014. 31 s., 12 s. příloh. Bakalářská práce. Vedoucí práce: Ing. Helena Škutková.

# PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma *Klasifikace organismů na základě nukleotidových četností* jsem vypracovala samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené práce dále prohlašuji, že v souvislosti s vytvořením této bakalářské práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědoma následků porušení ustanovení § 11 a následujících zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne ..... .....

(podpis autora)

# PODĚKOVÁNÍ

Děkuji vedoucímu bakalářské práce Ing. Heleně Škutkové za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé bakalářské práce a také za její vstřícnost a trpělivost.

V Brně dne ..... .....

(podpis autora)

# OBSAH

# SEZNAM OBRÁZKŮ

# ÚVOD

S objevem možností sekvenace DNA a její stále větší dostupnosti se započala zcela nová éra genetické analýzy a vyvstala i potřeba uchovávat všechna získaná data v jiné než papírové formě. Postupně začaly vznikat elektronické databáze sekvencí nukleotidů a proteinů, které spolu navázaly spolupráci, vzájemně si vyměňují informace, jsou veřejně přístupné a sdružují biomedicínská data získaná od laboratoří i jednotlivců z celého světa. Analýzy takového množství sekvencí již také nebylo reálné provádět ručně, a proto bylo nutné sestavit a naprogramovat automatizované algoritmy. Vznikaly algoritmy pro zarovnání sekvencí, grafickou reprezentaci jejich vzájemné blízkosti, matematické určování Euklidovské vzdálenosti a metody konstrukce fylogenetických stromů na základě evoluční příbuznosti. Bohužel tyto metody založené na zarovnaných sekvencích požadují velmi mnoho výpočetního prostoru, a proto je možné takto analyzovat pouze kratší úseky sekvencí. V současné době se směřuje k vývoji nových metod, které by se obešly bez nutnosti zarovnání analyzovaných sekvencí, byly tedy výpočetně méně náročné, ale přesto posloužily ke spolehlivé klasifikaci. Existující postupy založené na četnosti jednotlivých nukleotidů nebo jejich skupin v sekvenci jsou stále poměrně nové, ale již teď vykazují z hlediska účinnosti velmi slibné výsledky. Smyslem této práce je porovnat klasický přístup analýzy s moderními a to z hlediska účinnosti, přesnosti a výpočetní (časové) náročnosti.

# 1 BIOLOGICKÝ ÚVOD

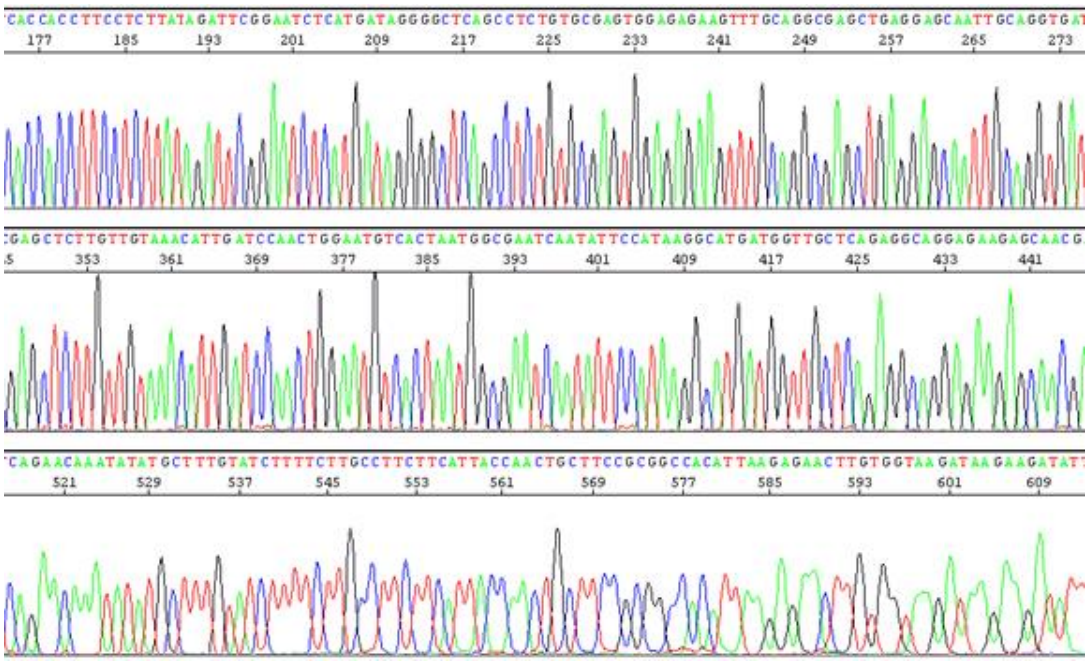
## 1.1 Fylogenetika

Fylogenetika je oborem zabývajícím se vývojovými vztahy mezi organismy na základě jejich evoluční příbuznosti a zkoumá, jaké změny provázely vývoj každého druhu či taxonu. Již před více než 100 lety se významní biologové jako například E. Haeckel snažili vysvětlit cestu od jednobuněčnosti k mnohobuněčným organismům pomocí několika hypotéz. Důležitou roli hrála jistá pozorovatelná analogie embryonálního růstu s jinými nižšími živočichy. Toto vedlo až k formulaci tzv. biogenetického zákona (E. Haecklem), který říká, že ontogeneze je vlastně zkrácenou fylogenezí. V současnosti je již však pokládán za překonaný. Z formulovaných hypotéz stojí za zmínku přinejmenším hypotéza koloniální, která považuje za prvopočátek kolonie jednobuněčných organismů. Samotná evoluce organismů je řízena diverzitou a mutací (změny DNA sekvencí). Fenotypová diverzita je zabezpečována genetickou variabilitou a to díky mnohonásobným alelám (varianty sekvencí na určitém místě chromozomu =lokusu) u mnoha genů. Mluvíme tedy o polymorfismu, existuje-li v populaci více alel. U mnoha eukaryotních organismů nalézáme diploidní sadu chromozomů. Můžeme tedy v případě, že má organismus na určitém lokusu dvě stejné alely říci, že je v tomto znaku homozygotní, v případě rozdílných alel heterozygotní. Polymorfní geny tedy u diploidních organismů poskytují prostor pro vznik velmi velkého množství genotypů. U člověka se odhaduje existence 23 500 genů, z nichž asi 1575 je heterozygotních. Teoreticky by tedy mohlo vzniknout $2^{1575} (\sim 10^{480})$ různých gamet. Toto číslo je tak vysoké, že jej nemůže být dosaženo ani za celou existenci lidstva. [1, 2] Specifickým podoborem fylogenetiky je neontologie, věda popisující současné organismy. K tomuto popisu využívá různé moderní srovnávací disciplíny z nichž nejlepší vypovídací hodnotu poskytuje srovnání genomů. Rozdíly a shody v genomech reprezentují skutečné příbuznosti i evoluční dobu diference těchto organizmů. [3].

## 1.2 Genom

Genomem můžeme označit všechny molekuly DNA (nebo RNA) nacházející se v živé soustavě, schopné replikace a přenosu na potomstvo. Dle jiné definice takto

můžeme nazvat souhrn všech genů v buňce. U prokaryot jsou geny uloženy těsně vedle sebe a obsahují minimum nekódujících sekvencí. Genom eukaryot lze rozdělit na strukturní geny, geny pro tRNA a rRNA, pseudogeny (podobnost se strukturními geny, ale nefunkční), regulační sekvence, pohyblivé sekvence (transpozóny, retropozóny) a mikrosatelity (nekódující repetitivně sekvence). Jen 28% sekvencí je transkribováno do RNA a pouze 1,1-1,4% genomu je přepisováno do proteinů. [4, 5] V 80. letech 20. Století byl navržen projekt lidského genomu. Očekávalo se, že tento projekt bude trvat 15 let a přijde na 3 miliardy dolarů. V roce 1981 se poprvé podařilo publikovat sekvenci DNA lidské mitochondrie o délce 16 569 pb. Existuje mnoho metod pro sekvenování DNA jako například SMRT (Single Molecule Real-Time), Maxam-Gilbertova a nejznámější Sangerova (Obrázek 1-1), která je založena na využití dideoxynukleotidů a následné elektroforéze. [6] Pro tuto sekvenaci jsou používány sekvence délky 200-300 pb a polyakrylamidový gel, který má malé póry a umožňuje tak rozlišit fragmenty lišící se od sebe i jen jedinou bází. [7, 8] Při následném porovnávání genomů byly objeveny blízké evoluční vztahy i mezi vzdálenými organismy. Díky kompletně osekvenovanému genomu *Drosophila Melanogaster* a přesným znalostem o rozložení genů se podařilo odhalit genetickou podstatu různých fyziologických dějů, například učení. Velmi překvapivý byl objev společného evolučního základu pro sluchové ústrojí i primární paměťové dráhy pro hmyz i vyšší živočichy (včetně člověka). Nečekané bylo také to, že počet genů naprosto neodpovídá složitosti organismu. Například u *Arabidopsis thaliana* (rostlina Huseníček rolní) byl zjištěn přibližně stejný počet genů jako u člověka. [7, 9]



Obrázek 1-1 Sangerova sekvenace [39]

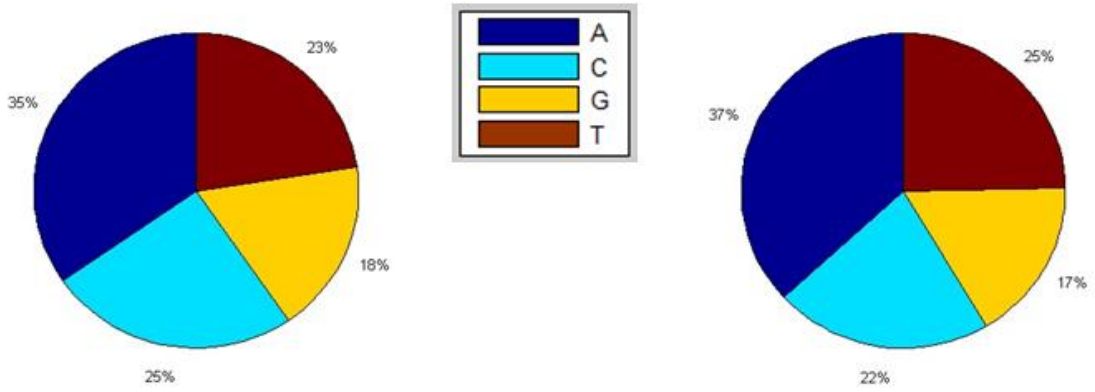
## 1.3 Význam nukleotidů

Deoxynukleová kyselina (DNA), nositelka genetické informace, se skládá z nukleotidů. Nukleotidy obsahují pentosu (cukr), zbytek kyseliny fosforečné a jednu ze čtyř dusíkatých bází A (Adenin),T (Thymin),C (Cytosin) nebo G (Guanin). Báze můžeme dělit dle jejich chemického složení na báze purinové (Adenin a Guanin) a pyrimidinové ( Cytosin a thymin) či podle obsahu volných radikálů (A,C -aminoskupina a G,T -ketoskupina). Díky komplementaritě bází se mohou báze mezi sebou párovat pouze jedním určitým způsobem a to A s T (2 vodíkové můstky) a C s G (3 vodíkové můstky) a tímto zaujmout energeticky nejvýhodnější konformaci. Samotná genetická informace je kódována pořadím nukleotidů (bází). [10] [11]Triplety bází (neboli kodony) jsou při translaci dekódovány a je podle nich syntetizován řetězec aminokyselin. Jelikož nukleotidy mohou obsahovat 4 různé báze a každý kodon obsahuje 3 báze, je celkový počet možných kodonů 64 (Tabulka 1-1). Mezi nimi nalezneme start kodony a stopkodony, které iniciují a ukončují translaci a 20 proteinogenních aminokyselin. [12] [13]

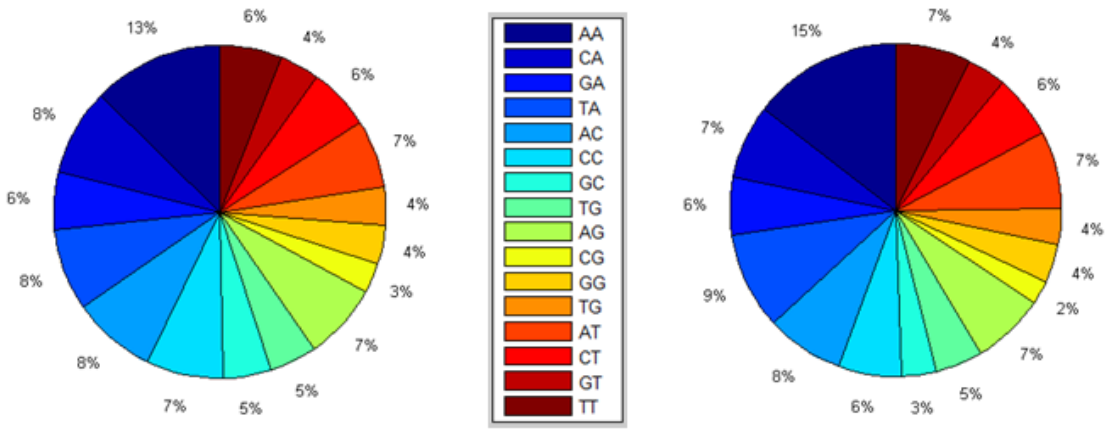
Tabulka 1-1 Všechny kombinace kodonů [40]

| | | Druhý nukleotid | | | | | |
|---|---|---|---|---|---|---|---|
| | | U | C | A | G | | |
| První nukleotid | U | UUU fenyalanin | UCU serin | UAU tyrosin | UGU cystein | U | Třetí nukleotid |
| | | UUC fenyalanin | UCC serin | UAC tyrosin | UGC cystein | C | |
| | | UUA leucin | UCA serin | UAA stop kodon | UGA stop kodon | A | |
| | | UUG leucin | UCG serin | UAG stop kodon | UGG tryptofan | G | |
| | C | CUU leucin | CCU prolin | CAU histidin | CGU arginin | U | |
| | | CUC leucin | CCC prolin | CAC histidin | CGC arginin | C | |
| | | CUA leucin | CCA prolin | CAA glutamin | CGA arginin | A | |
| | | CUG leucin | CCG prolin | CAG glutamin | CGG arginin | G | |
| | A | AUU isoleucin | ACU threonin | AAU asparagin | AGU serin | U | |
| | | AUC isoleucin | ACC threonin | AAC asparagin | AGC serin | C | |
| | | AUA isoleucin | ACA threonin | AAA lysin | AGA arginin | A | |
| | | AUG methionin | ACG threonin | AAG lysin | AGG arginin | G | |
| | G | GUU valin | GCU alanin | GAU kyselina asparagová | GGU glycin | U | |
| | | GUC valin | GCC alanin | GAC kyselina asparagová | GGC glycin | C | |
| | | GUA valin | GCA alanin | GAA kyselina glutamová | GGA glycin | A | |
| | | GUG valin | GCG alanin | GAG kyselina glutamová | GGG glycin | G | |

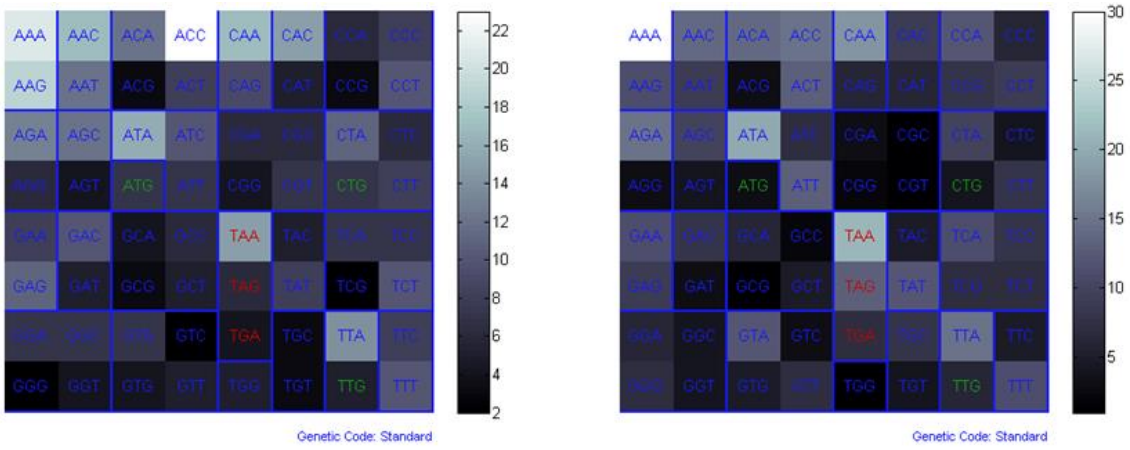
Genetická informace jako taková je degenerovaná, tzn., že jedna aminokyselina může být kódována pomocí více kodonů. [14] Pro klasifikaci organismů jsou využitelné nejen frekvence tripletů, ale i dimerů a jednotlivých bází (Graf 1-1), jelikož lze pozorovat rozdíly v jejich zastoupení u jednotlivých organismů. Ne všechny nukleotidy ale mají stejnou vypovídací hodnotu. Například počty Guaninu i Adeninu se v jednotlivých sekvencích liší minimálně, počty Cytosinu s Thyminem více. U dinukleotidů vedou počty těch, které obsahují Adenin, jelikož je to nejzastoupenější nukleotid (Graf 1-2). Rozdíly v počtech dinukleotidů jsou již lépe využitelné, už jen proto, že dávají vzniknout šestnácti třídám namísto čtyř. Teoreticky nejvýhodnější pro klasifikaci by mělo být porovnání četností tripletů (Graf 1-3), jelikož poskytují 64 tříd a jsou biologicky velmi významné, jelikož kódují syntézu aminokyselin. [15]




Graf 1-1 Procentuální zastoupení nukleotidů -vlevo *Papio papio*, vpravo *Eulemur fulvus*




Graf 1-2 Procentuální zastoupení dinukleotidů- vlevo *Papio papio*, vpravo *Eulemur fulvus*




Graf 1-3 Zastoupení tripletů -vlevo *Papio papio*, vpravo *Eulemur fulvus*

## 1.4 Mutace

Mutace se řadí k systematickým evolučním procesům a jejich výsledkem jsou pozměněné četnosti alel v populacích. Jedná se o všechny genetické změny, které mohou být přeneseny na potomstvo. Během mutace dochází ke změně genetické informace, která však stále dává smysl (i když pozměněný). Pokud jsou pravidla zápisu a kódování porušena, nejedná se o mutaci ale o poškození DNA. Rozlišujeme různé druhy mutací: bodové, řetězcové, chromozomové a genomové. Genové mutace probíhají na úrovni DNA vlákna a projevují se změněným pořadím nukleotidů. Může k nim dojít těmito mechanizmy

- Inzerce -Zařazení nadbytečných nukleotidů. Pokud jich není násobek tří, posune se čtecí rámec celé sekvence (tzv. frameshift mutation). Takto může dojít syntéze zcela jiného polypeptidu či předčasnému ukončení proteosyntézy.
- Delece- Ztráta jednoho nebo více nukleotidů z původní sekvence. Důsledky mohou být stejné jako při inzerci.
- Substituce- Záměna báze za jinou. Tyto mutace mohou, ale nemusí změnit výsledný produkt proteosyntézy. [16]

# 2 ANALÝZA PŘÍBUZENSKÝCH VZTAHŮ

## 2.1 Genetická informace v datech -genomika

Genomika je obor spadající pod genetiku a zabývající se studiem genomů organismů. Hlavním zájmem je získávání DNA sekvencí a jejich následná analýza ať už z pohledu jejich významu v organismu či evoluční souvislosti s jinými organismy. Sekvence DNA však obsahují kromě kódujících úseků (extrony) i úseky nenesoucí z genetického hlediska žádnou informaci. Tyto úseky nazýváme introny a jsou v průběhu transkripce vystřihávány pryč. Jejich přesný význam ani původ není dodnes zcela objasněn, existuje však několik teorií, jako například parazitický původ intronů. Kompletní sekvence všemožných organismů jsou dnes již veřejně dostupné v on-line verzi v mezinárodních databázích. [17] [18]

### 2.1.1 Databáze

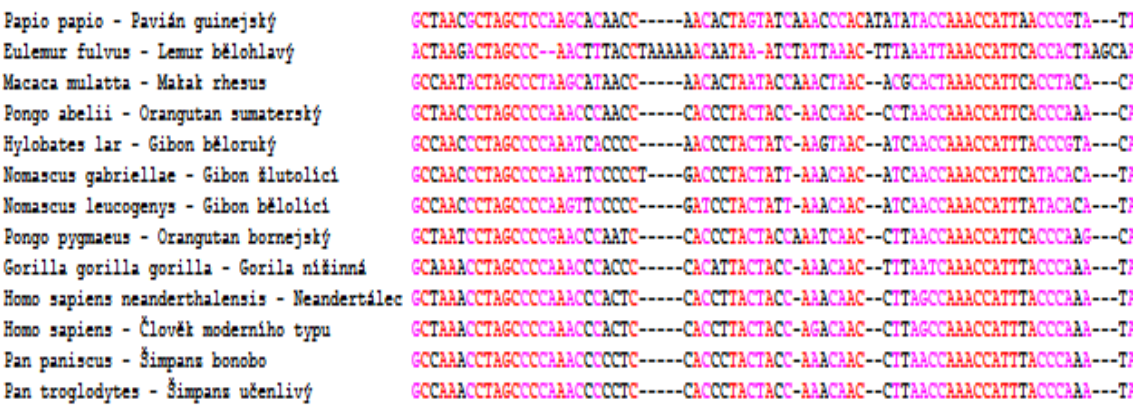
Již v 60. letech minulého století vznikla první sbírka sekvencí všech v té době známých proteinů. Byla vydána v tištěné podobě pod názvem Atlas of Protein Sequence and Structure a obsahovala i anotace. Se stále se zvětšujícím množstvím proteinových a dále i nukleotidových sekvencí vyvstala potřeba uchovávat tato data v elektronické podobě a mít možnost automaticky analyzovat jejich evoluční příbuznost. V roce 1982 tedy vzniká databáze DNA sekvencí při European Molecular Biology Laboratory (EMBL), dále GenBank při National Center for Biotechnology Information (NCBI) a po pár letech se přidává DNA Database of Japan (DDBJ). Od roku 1988 mezi sebou tyto databáze spolupracují, denně si vyměňují informace a formát jejich dat je standardizován. Po další době tuto trojici doplnila ještě SwissProt. Sekvence jsou kódovány dle standardu IUB nebo IUPAC (International Union of Biochemistry and Molecular Biology) a většinou uloženy ve formátu FASTA. Do databází jsou přijímána výsledná data ze všech uskutečněných sekvenací od institucí i jednotlivců po celém světě. Jelikož jsou tyto databáze primárním zdrojem sekvencí, mnoho databází je závislých na správnosti jejich dat. [19] [20]

### 2.1.2 Metody zarovnání sekvencí

Prvním krokem analýzy sekvencí je jejich vzájemné zarovnání. Obecně rozlišujeme dva přístupy; globální a lokální zarovnání. Needlemanův- Wunschův algoritmus provádí globální zarovnání, dává tedy přednost shodě celé délky zarovnaných sekvencí než dokonalé shodě v krátkých segmentech a je vhodnější pro podobnější sekvence. Naproti tomu algoritmus Smithův- Watermanův se stará o zarovnání lokální, ve kterém jsou hledány co nejpodobnější úseky jakýchkoliv délek a využití nachází především u zarovnání kratších sekvencí, které se od sebe poměrně liší ať už sekvenčně či délkově. Společná pro oba přístupy je snaha o nalezení co nejlepšího zarovnávacího score. Do samotného zarovnání vstupuje řada dalších parametrů jako penalizace mezer a váha páru (match) či nepáru (mismatch), pomocí nichž určíme, zda-li se vyplatí zavést na určitou pozici mezeru, nebo je výhodnější zde zarovnávání ukončit. Hodnoty pro jednotlivé shody i neshody získáme ze substituční (skórovací) matice. Pro mnohonásobné zarovnání (zarovnání více sekvencí) se využívá metoda sumy párů (která je však výpočetně velmi náročná), metoda spojování sousedů, či metoda CLUSTAL. Ta je v současné době dostupná ve třech verzích:

- CLUSTAL: všem sekvencím přiděluje stejnou váhu
- CLUSTALW: uživatel si sám volí váhy sekvencí a další parametry
- CLUSTALX: základní metoda avšak s genetickým rozhraním. [21] [22] [18]

V Matlabu se pro mnohočetné zarovnání využívá funkce multialign, která je součástí bioinformatického toolboxu. Můžeme si zde zvolit různé skórovací matice a další parametry. Na následujícím obrázku (Obrázek 2-1) je grafická ukázka výstupu zarovnání 13-ti sekvencí, které jsem si pro svou analýzu zvolila.



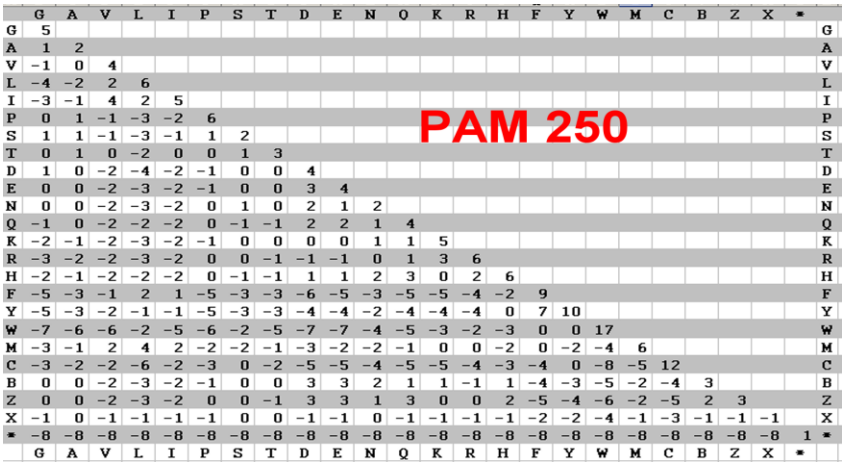

Obrázek 2-1 Mnohonásobné zarovnání 13 sekvencí primátů

### 2.1.3 Substituční matice (Substitution Matrix, Scoring Matrix)

Substituční matice je čtvercová matice, jejíž řádky a sloupce odpovídají znakům v sekvenci. Je souměrná podle hlavní diagonály, která odráží podobnost shodných symbolů (např. C-C). Hodnota u páru nesouhlasných znaků je shodná pro jakékoliv jejich pořadí (A-T má stejnou hodnotu jako T-A). Hodnota na průsečíku řádku a sloupce odpovídá příspěvku kombinace příslušných znaků k celkové podobnosti. Číselné hodnoty v praxi užívaných matic bývají stanoveny empiricky pozorováním skutečných sekvencí. Promítá se zde například frekvence záměn a výskytu konkrétních aminokyselin. Nejčastěji používané matice pro nukleotidy jsou PAM, NUC44, CLUSTALW a IUB (match=1,9; mismatch=0).

#### PAM (Point Accepted Mutation)

Byla odvozena v roce 1970 Margaret Dayhoffovou. Je sestavena na explicitním modelu z globálně zarovnaných sekvencí s podobností 85%. Udává pravděpodobnost mutace aminokyseliny (nukleotidu) do jiné (PAM1 =1 aminokyselina ze 100 zmutuje) v homologních sekvencích během evoluce. Matice vyšších hodnot jsou násobkem matice PAM1 s ní samotnou, nejvyšší je pak PAM250. [23]




Obrázek 2-2 Matice PAM 250 [33]

#### NUC44

Tato matice byla sestavena roku 1992 americkým profesorem Toddem Lowem. Nejnižší skóre obsažené v matici nabývá hodnoty -4 a nejvyšší hodnoty 5. [24]

## 2.2 Fylogenetický strom

Až do 50. let minulého století byly fylogenetické stromy, vyjadřující vzájemné příbuznosti organismů (či taxonů), sestavovány na základě subjektivních zkušeností jednotlivých odborníků. Tyto stromy vycházely z morfologických podobností. Dnešní metody však využívají podobnosti genetické, jelikož má daleko větší vypovídací hodnotu. Existuje mnoho metod konstrukce těchto stromů. Jedná se o metody distanční (pracují se vzdálenostmi) či znakové (vybírají nejvhodnější z více stromů).

Distanční metody:

- UPGMA (Unweighted Pair Group Method with Arithmetic mean): jedná se o shlukovou analýzu, která se snaží nalézt nejmenší hodnotu v distanční matici (=dvojice, která má k sobě nejblíže). Tyto příslušné taxonomické jednotky sloučí a určí jejich vzdálenost od všech ostatních. Takto se postupuje až do doby, kdy jsou všechny jednotky sloučeny v jednu. Poté tento postup znázorníme graficky ve formě stromu, jehož délka větví odpovídá evoluční vzdálenosti jednotlivých taxonů.
- Neighbour joining: na počátku je vytvořen jeden strom hvězdicovitého tvaru, kde všechny taxonomické jednotky odpovídají jednotlivým listům. V dalších krocích se strom shlukováním nejbližších jednotek rozkládá se snahou o co největší zmenšení celkové délky stromu. [25]

Znakové metody:

- Minimální evoluce: snaží se o minimalizaci součtu délky větví.
- Maximální parsimonie: hledá strukturu, která by odpovídala nejmenšímu počtu mutací při dosažení tohoto stavu.
- Maximum likelihood: ze všech možných stromů se vybere ten, s největší evoluční pravděpodobností vzniku. [26] [27]

# 3 NUMERICKÉ METODY ANALÝZY PŘÍBUZENSKÝCH VZTAHŮ

Po dlouhou dobu byly používány metody analýzy podobností DNA sekvencí založené na vzájemném zarovnání či grafickém znázornění (každou sekvenci reprezentuje křivka v Euklidovském prostoru). Vzhledem k jejich výpočetní náročnosti se objevily tendence vytvořit metody nezávislé na těchto postupech.

## 3.1 Numerická charakterizace DNA založená na dinukleotidech- Qi

Podstatou této metody je, že jednotlivé sekvence jsou reprezentovány pomocí matice nebo vektoru četností dinukleotidů. Nejdůležitějším rysem pak je, že nedochází k identifikaci pouze sousedních párů XY, ale i párů kde je X a Y odděleno jedním či více nukleotidy. Na rozdíl od ostatních metod je tímto zachována i informace kódovaná pořadím bází v DNA. Tato metoda je také velmi rychlá, jelikož nevyžaduje zarovnané sekvence ani jejich grafické znázornění a je vhodná pro analýzu jak krátkých tak i dlouhých sekvencí DNA. Jelikož existují čtyři různé nukleotidové báze, pracuje tato metoda s 16-ti rozměrným vektorem, jež obsahuje všechny možné kombinace dinukleotidů, kde X a Y jsou přímo vedle sebe a dalšími přídavnými vektory, v nich jsou uloženy hodnoty četností X a Y, které od sebe dělí jeden, dva, či více nukleotidů. Speciálními postupy pro tvorbu distanční matice na základě těchto četností jsou City blok distance $d_1(s, h)$ (3-1), který porovnává rozdíly mezi maticemi četností $F(s)$ a $F(h)$ jako

$$\mathrm{d}_1(\mathrm{s},\mathrm{h}) = \sum_{1\le \mathrm{i},1\le \mathrm{j}\le 16} \left|\mathrm{F}_{\mathrm{ij}}(\mathrm{s}) - \mathrm{F}_{\mathrm{ij}}(\mathrm{h})\right| \tag{3-1}$$

a Kosinové distance $d_2(s,h)$ (3-2), který k tomuto využívá úhlu mezi vektory a funkce cosinus

$$\mathrm{d}_2(\mathrm{s},\mathrm{h}) = 1 - \cos\left(\hat{F}(\mathrm{s}), \hat{F}(\mathrm{h})\right), \tag{3-2}$$

kde $\hat{F}(s)$ *a* $\hat{F}(h)$ je soubor všech vektorů frekvencí dinukleotidů dané sekvence. [15]

## 3.2 Numerická charakterizace DNA pomocí četností slov různých délek- Yang

Tato metoda využívá k analýze jednotlivých sekvencí vektor četností vyskytujících se charakteristických slov o různé délce. Četnosti v tomto vektoru jsou poté seřazeny vzestupně a výsledný vektor je získán jako ukazatele pořadí výskytů jednotlivých slov. Ze čtyř znaků vyskytujících se v sekvenci: A, C, G, T, můžeme vytvořit $n = 4^k$ možných slov délky $k$. Četnosti výskytu slova $i = \{1, 2, 3, \dots, n\}$ délky $k$ jsou označeny jako $c(w_{k,i})$. Celá sekvence je potom reprezentována vektorem četností:

$$C_k = \left(c\left(w_{k,1}\right), c\left(w_{k,2}\right), c\left(w_{k,3}\right), \dots, c\left(w_{k,n}\right)\right). \tag{3-3}$$

Seřazený vektor $S_k$ četností z vektoru $C_k$ ((3-3) vzestupně označíme jako:

$$S_k = \left(c\left(w_{k,r_1}\right), c\left(w_{k,r_2}\right), c\left(w_{k,r_3}\right), \dots, c\left(w_{k,r_n}\right)\right). \tag{3-4}$$

Vektor ukazatelů výskytů slov $O_k$ definujeme:

$$O_k = \left(o\left(w_{k,1}\right), o\left(w_{k,2}\right), o\left(w_{k,3}\right), \dots, o\left(w_{k,n}\right)\right). \tag{3-5}$$

K určení výsledné vzdálenosti dvou sekvencí použijeme Euklidovskou vzdálenost vektorů $O_k$:

$$d_k(s_1, s_2) = \sqrt{\sum_{i=1}^{4^k} (o^{s_1}\left(w_{k,i}\right) - o^{s_2}\left(w_{k,i}\right))^2}. \tag{3-6}$$

Vytváříme-li konsensuální strom pro více slov různé délky $k$, distanci normalizujeme celkovým počtem slov, tzn. podělíme hodnotou $4^k$. [28]

## 3.3 Charakterizace DNA pomocí tripletů

Jedná se o alternativní přístup k porovnání DNA sekvencí, kde se místo Levenshteinova porovnávání řetězce používají invarianty v sekvenci DNA, triplety. Jelikož se opět vychází ze sekvencí DNA obsahujících čtyři nukleotidy (A,C,G,T),

používá se pro uložení četností tripletů kubická matice 4x4x4; tedy celkem 64 hodnot a zkoumají se rozdíly ve frekvencích výskytu tripletů. Podobnost těchto matic se opět zkoumá většinou pomocí Euklidovy vzdálenosti či skalárního součinu nebo můžeme tyto dvě metody kombinovat. Tato metoda není ještě zcela ideálně vyvinutá, spíše se snaží nastínit možné směry, kterými se může klasifikace pomocí četností tripletů ubírat. Autoři spíše vznáší otázky, co by bylo nejlepší, jestli používat matice, kde každá položka odpovídá konkrétnímu tripletu nebo tzv. kondenzované matice, ve kterých se každá hodnota vztahuje k více tripletům. Zda by mělo existovat pořadí preferencí pro volbu charakterizace pomocí četností nukleotidů, dinukleotidů nebo tripletů. Jestli v případě využití tripletů má být analýza založena na nepřekrývajících se trojicích nebo všech možných trojicích v sekvenci a další... [29]

## 3.4 Numerická charakterizace pomocí dinukleotidů s využitím jejich chemických vlastností

Nukleotidy můžeme dělit dle jejich chemických vlastností do tří skupin s dvěma variantami v každé z nich (purinové/ pyrimidinové, slabá/ silná H vazba, keto/ amino skupina; pracovat tedy budeme s šesti znakovým vektorem). Těchto vlastností můžeme dobře využít při numerické klasifikaci DNA. Tento přístup vykazuje mnohem lepší výsledky než přístupy čistě matematické, jelikož jsou k porovnání použity vždy dvojice nukleotidů stejného druhu a mající tedy pro sekvenci stejný význam. K tvorbě samotných distančních matic můžeme použít 3 postupy

- Euklidovská vzdálenost, která se počítá vždy pro příslušné významové dvojice . Mějme tedy sekvenci s počtem *N* bází $S= s_1 s_2 \ldots s_N$, $s_i \in \{A,C,G,T\}$. Euklidovskou vzdálenost například pro třídu purin/pyrimidin vypočteme takto (3-7) : (i,j) prvek matice AG je definován jako

$$[AG]_{ij} = \sqrt{(A_{ij} - mG_{ij})^2 + (T_{ij} - mC_{ij})^2}. \tag{3-7}$$

- Matice distancí „cesty“ *PD*, která vychází z klasické Euklidovské distance, ale diference se počítá vždy ze dvou po sobě jdoucích prvků takto ((3-8).

$$[PD]_{ji} = [PD]_{ij} = [ED]_{i,i+1} + [ED]_{i+1,i+2} + \cdots + [ED]_{j-1,j} \quad (3\text{-}8)$$

$$i < j, [PD]_{ii} = 0,$$

kde *ED* je jedna z matic Euklidovských distancí.

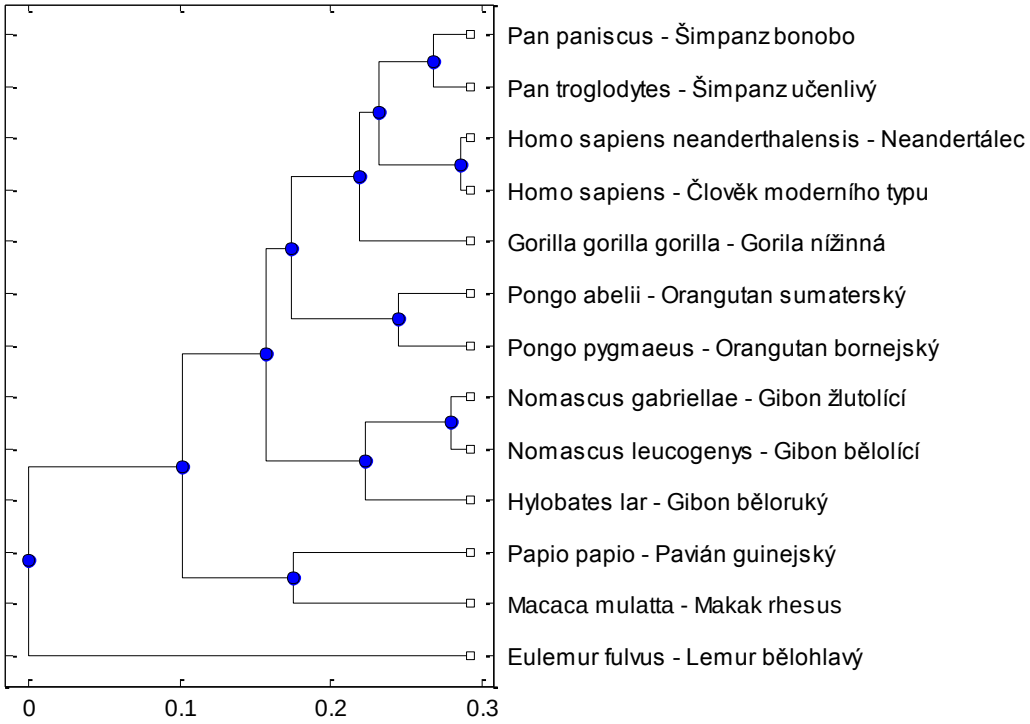
- Matice podílů *E/P*, definovaná jako podíl korespondujících prvků z *ED* a *PD*:

$$[E/P]_{ij} = [ED]_{ij}/[PD]_{ij}, i \neq j, [EP]_{ii} = 0. \quad (3\text{-}9)$$

Jelikož zpracováváme nezarovnané sekvence, vektory četností ještě normalizujeme pomocí podělení vektorů četností délkou sekvence. [30]

## 3.5 Zpracování nových numerických metod

Rozhodla jsem se parafrázovat výše zmíněné numerické metody a zkusit vytvořit zjednodušenou verzi klasifikace organismů pouze na základě nukleotidových, dinukleotidových a tripletových četností a následnou konstrukcí fylogenetického stromu porovnat jejich účinnost vzhledem k metodám klasickým s nutností zarovnání (Obrázek 3-1). Tyto metody neberou v potaz oblast výskytu daných nukleotidů, ale pouze jejich počet, což je činí pravděpodobně méně přesnými. Ke zjištění jednotlivých četností budu využívat funkce z Matlabu: *basecount*, *dimercount* a *codoncount*. Hodnoty do matice distancí získám výpočtem normalizované hodnoty podobnosti pomocí naprogramovaného algoritmu (viz *matice_distanci*) či využitím některé z metod dostupných v bioinformatickém toolboxu Matlabu . Výsledný fylogenetický strom zobrazuji metodou UPGMA pomocí funkce *seqlinkage*. K analýze jsem použila 13 sekvencí DNA primátů (Tabulka 4-1).




Obrázek 3-1 Fylogenetický strom pomocí klasického přístupu s vícenásobným zarovnáním

## Klasifikace pomocí četností nukleotidů

Nejprve zjistíme četnosti $\mu$ jednotlivých nukleotidů v sekvencích. Jednotlivé hodnoty matice distancí $B$ (3-10),

$$B(i,j) = \sqrt{(\mu_{i1}-\mu_{j1})^2 + +(\mu_{i2}-\mu_{j2})^2 \dots (\mu_{iN}-\mu_{jN})^2}\,, pro\ N = 1,2, \dots 4 \tag{3-10}$$

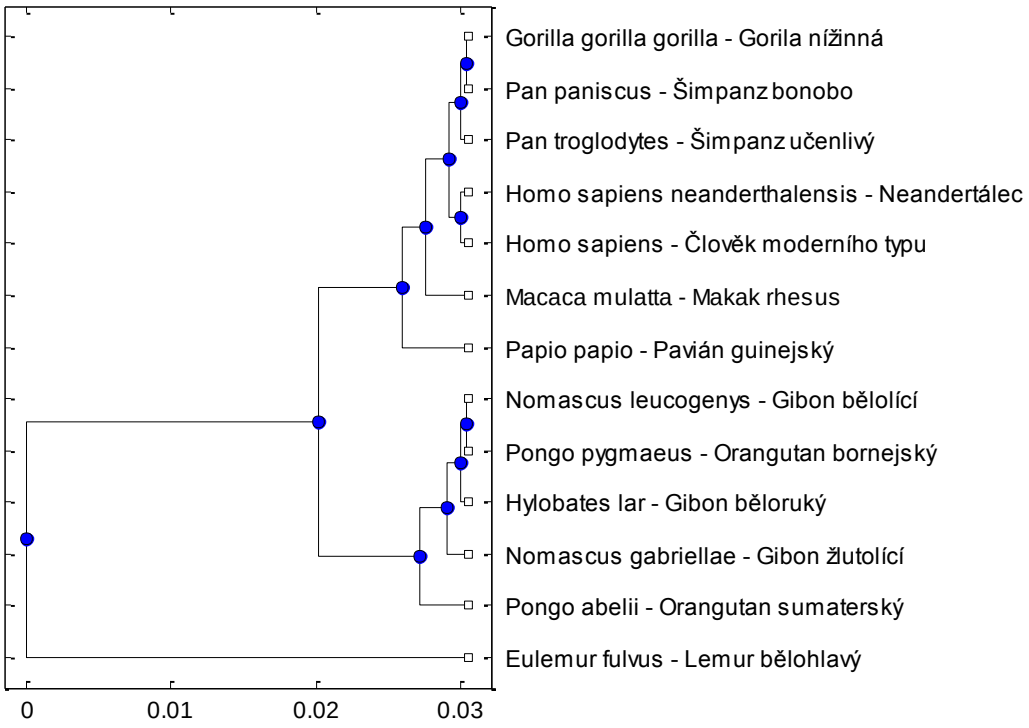
poté získáme pomocí Euklidovské vzdálenosti, kde $N$ je počet všech možných nukleotidů (tedy A, C, G, T). Jelikož však pracujeme s nezarovnanými sekvencemi, je nutné nejprve použít standardizaci tím, že jednotlivé četnosti podělíme délkou sekvence.

## Klasifikace pomocí četností dinukleotidů

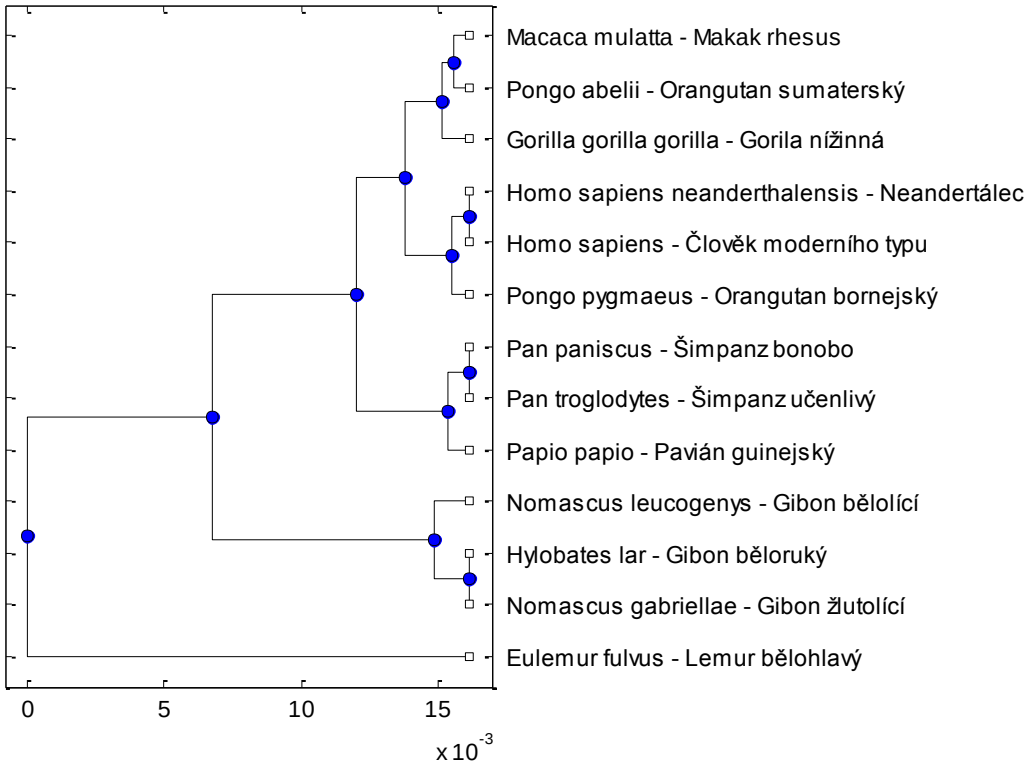
Tento postup vychází z klasifikace pomocí četností nukleotidů. Výpočet hodnot pro matici distancí provedeme obdobně jako u předešlého postupu (3-10), avšak s tím rozdílem, že $N$ nyní nabývá hodnot (1,2,…16), dle všech možných kombinací dinukleotidů.

## Klasifikace pomocí četností tripletů

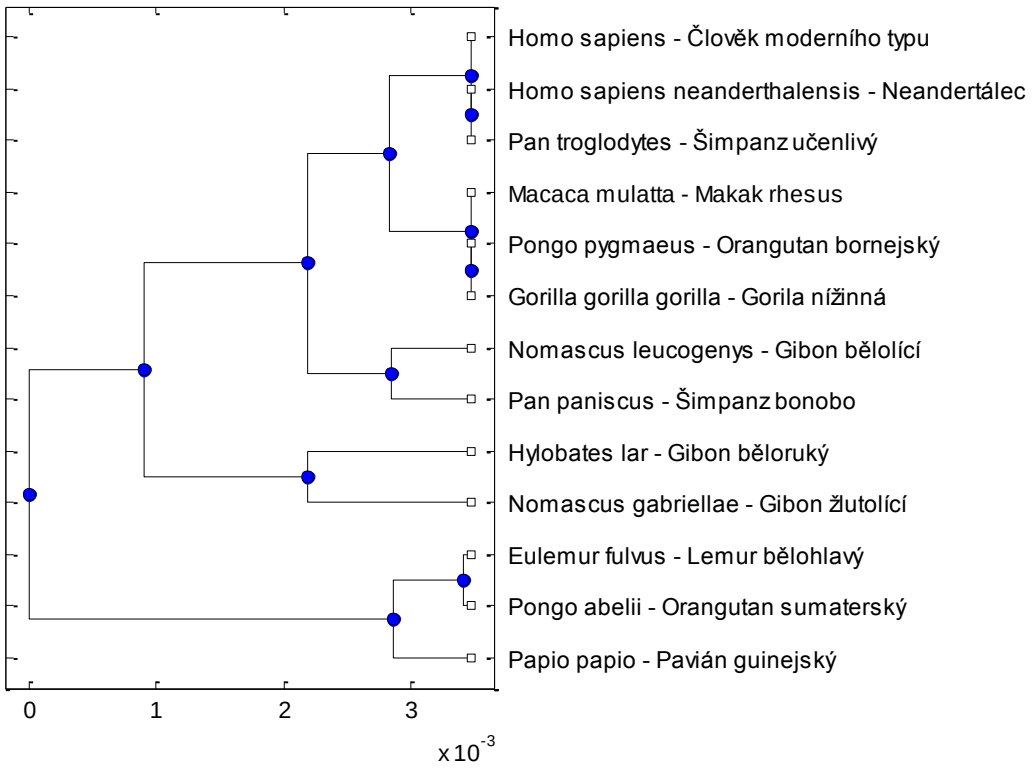
Zopakujeme výše uvedený postup výpočtu matice distancí (3-10). Nyní však existuje dokonce 64 možných tripletů a proto $N$ nyní nabývá hodnot (1,2,…64).




Obrázek 3-2Fylogenetický strom pomocí četností nukleotidů




Obrázek 3-3 Fylogenetický strom pomocí četností dinukleotidů




Obrázek 3-4 Fylogenetický strom pomocí četností tripletů

# 4 PROGRAM PRO TVORBU FYLOGENETICKÝCH STROMŮ DLE RŮZNÝCH METOD ANALÝZY DNA

Jako rozhraní pro tvorbu programu, vytvářejícího fylogenetické stromy a obsahujícího grafické uživatelské rozhraní jsem zvolila Matlab. Program je tvořen nově sestavenými funkcemi (viz Zdrojové kódy) a zároveň využívá i již definované funkce z bioinformatického toolboxu Matlabu, jako například *pdist* (pro výpočet distancí) *seqlinkage* (pro vytvoření stromu).

## 4.1 Výběr vhodných sekvencí

Výběr vhodné testovací sekvence je ovlivněn mnoha faktory. Nesmí být příliš dlouhá, kvůli výpočetní náročnosti. Měl by se vyskytovat její ekvivalent u co největšího počtu dalších organismů, aby byla testovací základna dostatečně velká. Vybraná sekvence by také měla být dostatečně charakteristická pro každý organismus.

Pro svou analýzu jsem zvolila sekvence 16S mitochondriální rRNA třinácti různých druhů primátů uvedených v tabulce (Tabulka 4-1). Buňky více-buněcných živočichů obsahují mitochondriální DNA v kružnicové formě a každý živočišný druh má svůj vlastní typ mtDNA, což činí tyto sekvence velmi vhodné ke všem různým DNA analýzám příbuzosti. Mitochondriální sekvence také patří k nejčastěji sekvenovaným a bývají tedy spolehlivé. Délka jednotlivých sekvencí se pohybuje od 1557 do 1575 bp.

Tabulka 4-1 Použité sekvence primátů- 16S mitochondriální rRNA

| Organismus | Délka sekvence |
|---|---|
| *Papio papio* - Pavián guinejský | 1562 bp |
| *Eulemur fulvus* - Lemur bělohlavý | 1575 bp |
| *Macaca mulatta* - Makak rhesus | 1558 bp |
| *Pongo abelii* - Orangutan sumaterský | 1560 bp |
| *Hylobates lar* - Gibon běloruký | 1558 bp |
| *Nomascus gabriellae* - Gibon žlutolící | 1558 bp |
| *Nomascus leucogenys* - Gibon bělolící | 1557 bp |
| *Pongo pygmaeus* - Orangutan bornejský | 1558 bp |
| *Gorilla gorilla gorilla* - Gorila nížinná | 1558 bp |
| *Homo sapiens neanderthalensis* - Neandertálec | 1558 bp |
| *Homo sapiens* - Člověk moderního typu | 1559 bp |
| *Pan paniscus* - Šimpanz bonobo | 1559 bp |
| *Pan troglodytes* - Šimpanz učenlivý | 1558 bp |

Jako druhou analyzovanou skupinu jsem zvolila 30 druhově rozmanitých sekvencí kompletní mitochondriální DNA savců. Nalezneme v ní organismy od člověka, přes myš až po plejtváka (viz Tabulka 4-2). Délka jednotlivých sekvencí se v tomto případě pohybuje od 16295 do 17245 bp, rozdíl délek tedy činí až 950 bp.

Tabulka 4-2 Použité sekvence kompletní mitochondriální DNA savců

| **Organismus** | **Délka sekvence** | **identifikátor** |
|---|---|---|
| *Homo sapiens*- Člověk moderního typu | 16569 bp | V00662.1 |
| *Pan paniscus*- Šimpanz bonobo | 16563 bp | D38116.1 |
| *Pan troglodytes*- Šimpanz učenlivý | 16554 bp | D38113.1 |
| *Gorilla gorila*- Gorila nížinná | 16364 bp | D38114.1 |
| *Pongo pygmaeus*- Orangutan bornejský | 16389 bp | D38115.1 |
| *Hylobates lar*- Gibon běloruký | 16472 bp | X99256.1 |
| *Papio hamadryas*- Pavián pláštíkový | 16521 bp | Y18001.1 |
| *Equus caballus*- Kůň domácí | 16660 bp | X79547.1 |
| *Ceratotherium simum*- Nosorožec tuponosý | 16832 bp | Y07726.1 |
| *Phoca vitulina*- Tuleň obecný | 16826 bp | X63726.1 |
| *Halichoerus grypus*- Tuleň kuželozubý | 16797 bp | X72004.1 |
| *Felis catus*- Kočka domácí | 17009 bp | U20753.1 |
| *Balaenoptera physalys*- Plejtvák myšok | 16398 bp | X61145.1 |
| *Balaenoptera musculus*- Plejtvák obrovský | 16402 bp | X72204.1 |
| *Bos taurus*- Tur domácí | 16338 bp | V00654.1 |
| *Rattus norvegicus*- Potkan obecný | 16300 bp | X14848.1 |
| *Mus musculus*- Myš domácí | 16295 bp | V00711.1 |
| *Didelphis virginiana*- Vačice virginská | 17084 bp | Z29573.1 |
| *Macropus robustus*- Klokan horský | 16896 bp | Y10524.1 |
| *Ornithorhynchus anatinus*- Ptakopysk podivný | 17019 bp | X83427.1 |
| *Sciurus vulgaris*- Veverka obecná | 16507 bp | AJ238588.1 |
| *Myoxus glis*- Plch obecný | 16602 bp | AJ001562.1 |
| *Cavia porcellus*- Morče domácí | 16801 bp | AJ222767.1 |
| *Equus asinus*- Osel domácí | 16670 bp | X97337.1 |
| *Rhinoceros unicornis*- Nosorožec indický | 16829 bp | X97336.1 |
| *Canis familiaris*- Pes domácí | 16727 bp | U96639.2 |
| *Ovis aries*- Ovce domácí | 16616 bp | AF010406.1 |
| *Sus scrofa*- Prase divoké | 16680 bp | AJ002189.1 |
| *Hippopotamus amphibius*-Hroch obojživelný | 16407 bp | AJ010957.1 |
| *Oryctolagus cuniculus*- Králík divoký | 17245 bp | AJ001588.1 |

## 4.2 Použité metody

V programu jsem pro charakterizaci sekvencí DNA použila několik metod. Jako srovnávací etalon jsem zvolila klasický přístup s využitím vícenásobného zarovnání sekvencí a výpočtem distancí pomocí metody Jukes-Cantor. Jako zástupce nového přístupu jsem zvolila tři numerické metody- Qi (viz 3.1), Yang (viz 3.2) a jednoduchou klasifikaci pouze na základě četností dinukleotidů (viz 3.5). U všech metod je pro tvorbu výsledného fylogenetického stromu použita průměrovací metoda UPGMA. Cílem je vytvořit fylogenetické stromy pomocí různých přístupů k reprezentaci DNA a provést porovnání výsledků s přístupem klasickým, obsahujícím vícenásobné zarovnání. Jelikož se jako hlavní výhoda numerických metod uvádí nepotřebnost zarovnání sekvencí a tím i značné snížení výpočetních nároků, doporučuji v programu nastavení parametrů u metod Qi a Yang omezit na malý interval hodnot. U metody Qi je však přesto možno použít maximální možné hodnoty parametrů odpovídající vzorci

$$p \leq \frac{n-1}{2}, \tag{4-1}$$

kde $p$ vypovídá o vzájemné vzdálenosti dvojice nukleotidů v sekvenci $s$ o délce $n$ a jejich nastavení je tedy ponecháno čistě na uživateli. Výsledný strom je sestaven jako průměrový konsenzus všech stromů o různých distancích mezi nukleotidy. Co se týče metody Yang, přistoupila jsem k omezení na nejvyšší délku slova osm, jelikož při tvorbě výsledného fylogenetického stromu je vždy použit průměrový konsensus všech stromů o délce slov dva až nastavené maximum, tato hodnota se jeví jako ideální kompromis přesnosti a výpočetní náročnosti. Celkový počet slov je tak v tomto případě 87 376. Samotní autoři této metody však na větší časovou náročnost upozorňují. [28]

## 4.3 Typy konsensuálních stromů

**Striktně konsenzuální strom:**

Zobrazí pouze ta binární větvení, která se vyskytují ve všech dílčích stromech. V případě, že si binární větvení odporují, jsou nahrazena větvením vyššího řádu.

**Většinově konsenzuální strom:**

Zobrazuje větvení, která mají v jednotlivých dílčích stromech největší poměrné zastoupení.

**Průměrový konsenzuální strom:**

Tento strom pracuje s prostým aritmetickým průměrem daných větví.

## 4.4 Hodnocení podobnosti stromů

K vyhodnocení vzájemné podobnosti fylogenetických stromů se využívá několik metod. Uveďme si tedy alespoň dvě z nich:

### 4.4.1 Pearsonův korelační koeficient

Korelační koeficient slouží ke stanovení typu a síly závislosti mezi dvěma veličinami -v našem případě mezi vypočtenými distancemi mezi sekvencemi. Nejjednodušším vztahem dvou metrických proměnných je vztah lineární. Tento koeficient nabývá hodnot v intervalu (-1; 1), kde hranice -1 označuje nepřímou lineární závislost a hranice 1 přímou lineární závislost. [31]

### 4.4.2 Robinson- Fouldova vzdálenost

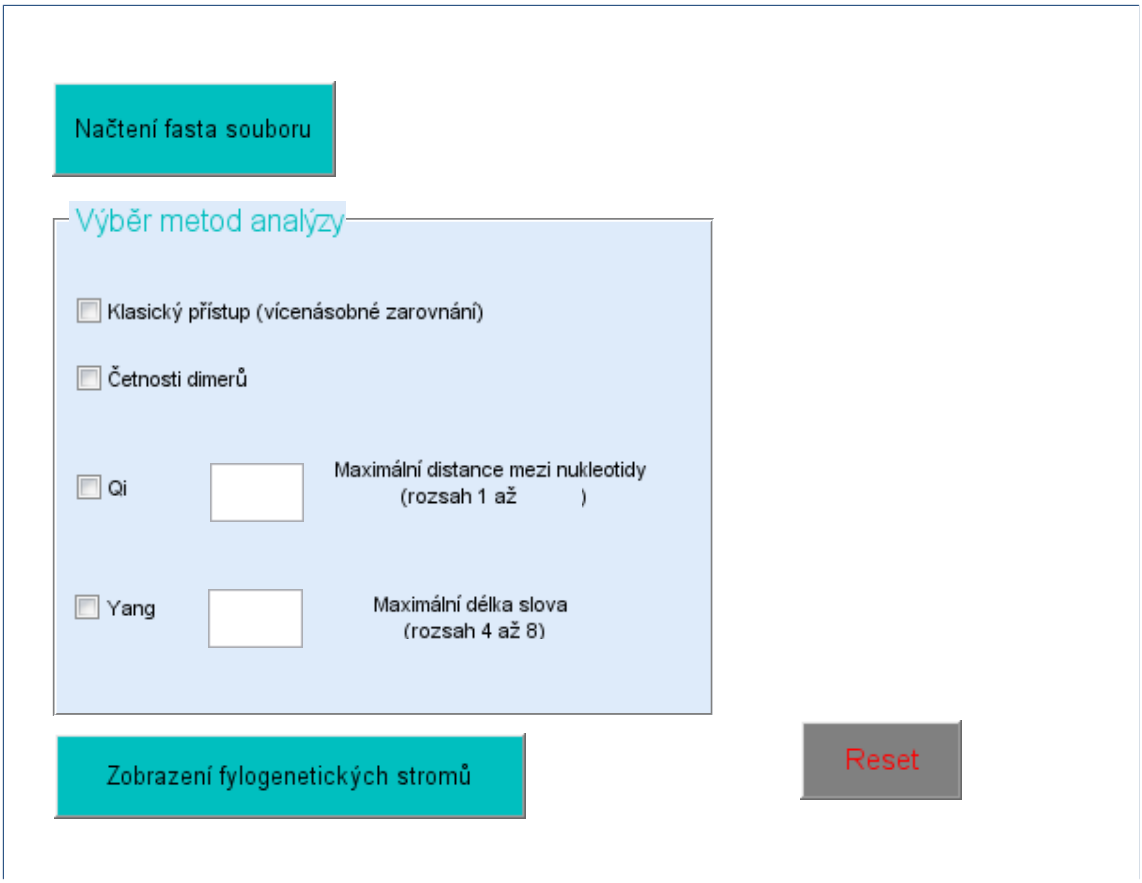
Základem je obsahové porovnání shluků dvou stromů. Postupně takto můžeme vyhodnotit nejpřesnější strom vůči zvolenému referenčnímu. Robinsonovu -Fouldovu vzdálenost (dále jen R-F vzdálenost) $d_{r,s}$ získáme ze vzorce

$$d_{r,f} = \frac{(n_{c1-2} + n_{c2-1})}{2 \cdot n}, \tag{4-2}$$

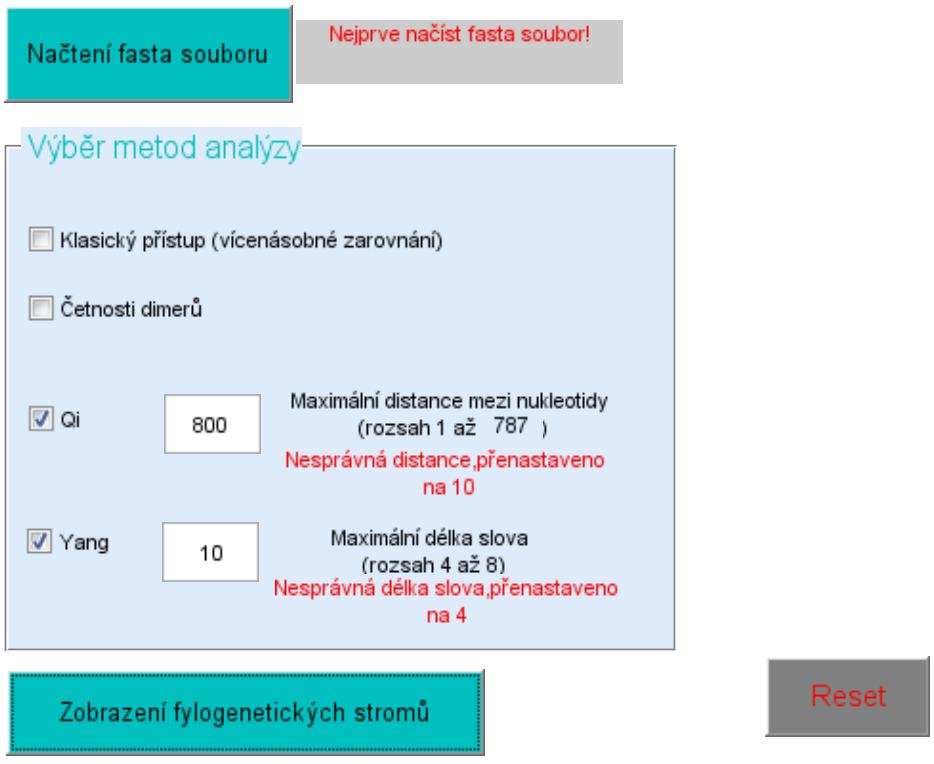
kde $n_{c1-2}$ vyjadřuje počet shluků vyskytujících se pouze v prvním stromu, $n_{c1-2}$ počet shluků vyskytujících se pouze v druhém stromu a $n$ počet shluků u obou sekvencí. [32]

## 4.5 Grafické uživatelské rozhraní

Grafické uživatelské rozhraní jsem volila jednoduché, uživatelsky příjemné s intuitivním ovládáním. Vstupní obrazovka obsahuje tlačítka pro načtení fasta souboru a vykreslení výsledných stromů a dále nabídku čtyř různých metod analýzy DNA včetně nastavení jejich parametrů (u metod Qi, Yang) (viz Obrázek 4-2). Maximální možná distance mezi nukleotidy u metody Qi je vypočítána vždy s ohledem na délky aktuálně načtených sekvencí. Pro načtení fasta souboru se sekvencemi jsem použila klasické vyskakující dialogové okno, kde uživatel nastaví cestu až k místu úložiště souboru. Pokud tak neučiní a přesto spustí vykreslování stromů, zobrazí se chybová hláška (viz Obrázek 4-1). Stejně tak pokud uživatel zadá parametry některé z metod mimo povolené intervaly, upozorní ho chybová hláška (viz Obrázek 4-1), parametr se automaticky přenastaví (o čemž bude informován) a analýza proběhne. Pro zrušení všech učiněných nastavení slouží tlačítko „reset“.



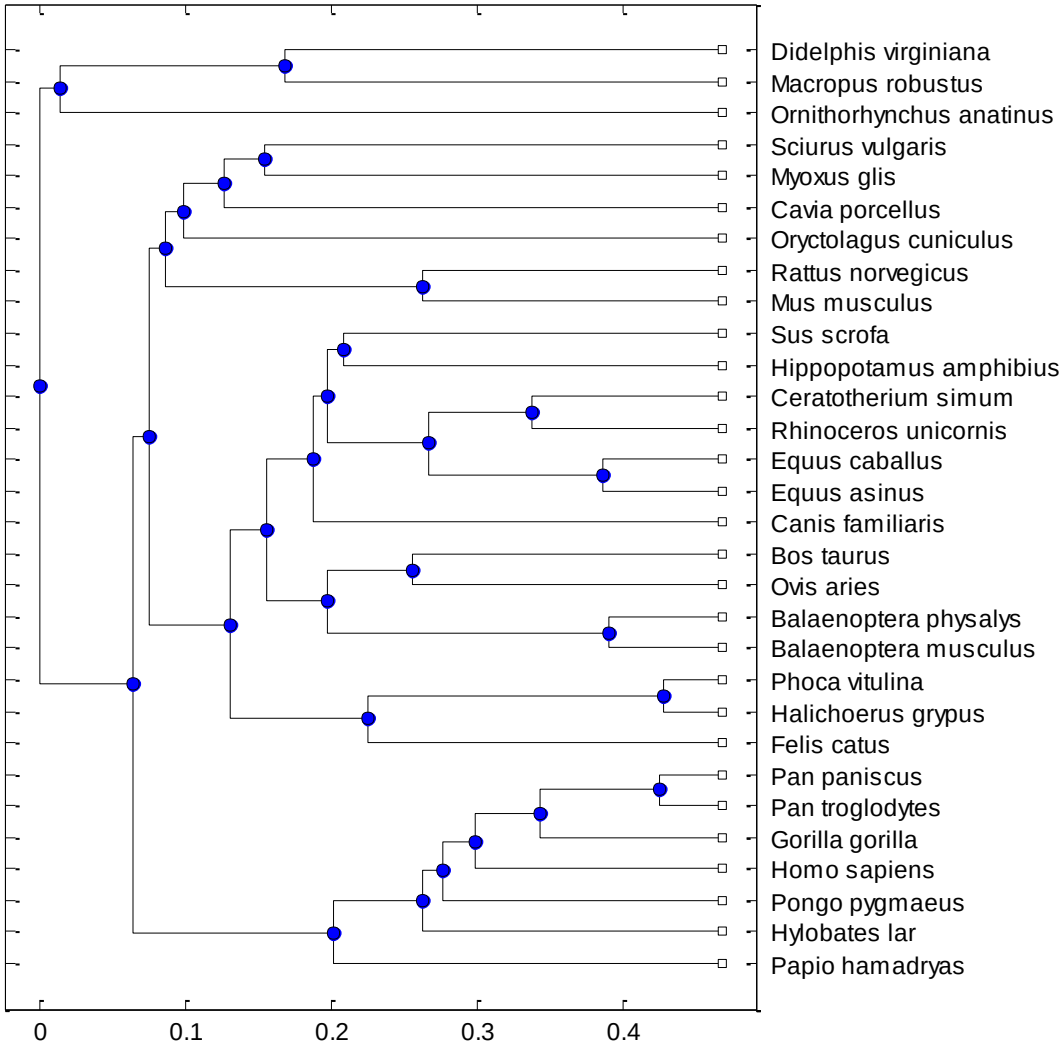

Obrázek 4-2 Vstupní obrazovka grafického rozhraní




Obrázek 4-1 Vstupní obrazovka grafického rozhraní s chybovými hláškami

## 4.6 Úspěšnost metod

Úspěšnost daných metod byla stanovena pomocí R-F vzdálenosti (viz 4.4.2), kde jako srovnávací posloužil fylogenetický strom vytvořený pomocí klasického přístupu s využitím vícenásobného zarovnání. Ukázku tohoto stromu pro skupinu savců představuje Obrázek 4-3. Analýza skupiny primátů přinesla zřetelně přesnější výsledky u všech použitých metod. Důvodem je jak délka jednotlivých sekvencí, které jsou cca 10 krát kratší a i variabilita délek je pouze několik nukleotidů na rozdíl od skupiny savců, kde rozdíl délek činí až 950.




Obrázek 4-3 Fylogenetický strom pomocí klasického přístupu, sekvence savců

### Četnosti dimerů

Využití pouze četností dimerů ke klasifikaci se (dle předpokladu) ukázalo nejméně přesné. U skupiny primátů R-F vzdálenost nabývá hodnoty 0,625, u savců dokonce 0,9. U fylogenetického stromu primátů se podařilo identifikovat dvojici *Homo sapiens, Homo sapiens neanderthalensis* i *Pan troglodytes, Pan paniscus* a dále i celou skupinu tří gibonů (*Hylobates lar, Nomascus leucogenys* a *Nomascus gabriellae*) i když s mírnou nepřesností. Podařilo se i správně určit vzdálenou příbuznost *Eulemura fulvus* ke všem ostatním. U savců byla správně zařazena pouze dvojice nosorožců a tuleňů.

### Qi

Metoda Qi byla vyhodnocena jako druhá nejpřesnější, přičemž se se zvyšováním maximální distance mezi nukleotidy nejprve zvyšuje i přesnost metody, a to u primátů z R-F vzdálenosti 0,3 (při maximální distanci 1) až na 0,083 (při maximální distanci 100). Další zvyšování však již zapříčinilo naopak větší nepřesnost a výpočetní nároky se již při nastavení maximální distance na polovinu povolené hodnoty plynoucí ze vzorce (4-1) staly nepřípustnými. Při tvorbě stromu pro skupinu savců se výpočetní náročnost metody zřetelně projevila již při nastavení maximální distance na hodnotu 20 a její přesnost až po tuto hodnotu stagnovala. Na stromech byly přesto pozorovatelné drobné změny v uspořádání. Nejlepší výsledky metoda poskytla při maximální distanci 30 a 50.

### Yang

Fylogenetické stromy vytvořené pomocí této metody poskytují velmi uspokojivé výsledky a to již při použití slov o maximální délce 4. Na přesnost shodnou s metodou Qi u primátů se dostaneme již, když slova obsahují nejvýše 6 znaků. Přitom k tomu potřebujeme několikanásobně méně výpočetního prostoru (času). Při klasifikaci skupiny savců je rozdíl v účinnosti obou metod ještě markantnější, jelikož metoda Yang začíná při maximální délce slova 4 na R-F vzdálenosti 0,59, kdežto metoda Qi této hodnoty nedosahuje ani při zvýšení maximální distance mezi nukleotidy na 50. Obecně se potvrdilo zlepšení účinnosti s použitím větší maximální délky slova.

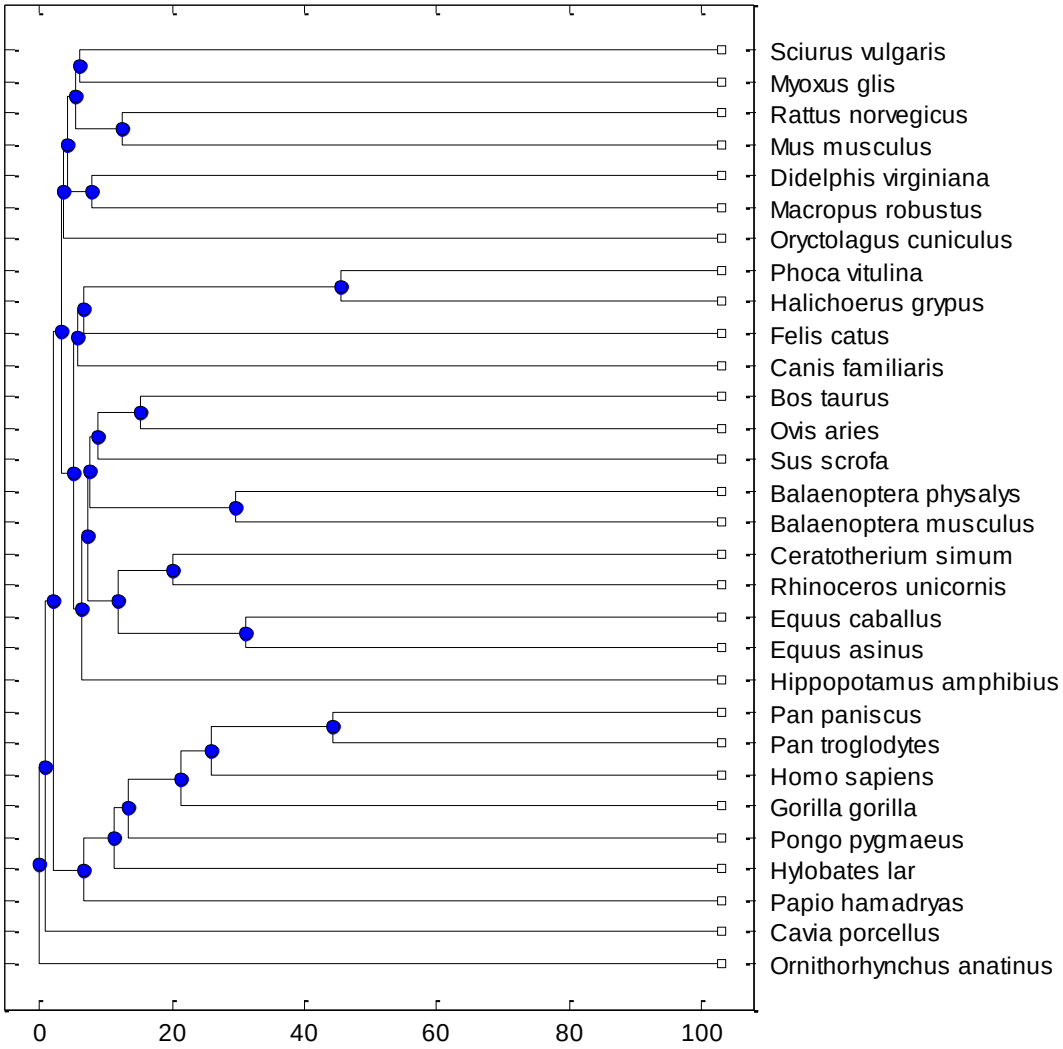
Výsledné R-F vzdálenosti pro všechny použité metody i obě skupiny sekvencí jsou obsaženy v následujících tabulkách:

Tabulka 4-3 Výsledné R-F vzdálenosti sekvencí primátů

| **Primáti- 16S mitochondriální rRNA** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Četnost dimerů | $d_{r,f}$ | 0,625 | | | | | | |
| | | | | | | | | |
| Qi | Distance | 1 | 10 | 40 | 70 | 100 | 250 | 393 |
| | $d_{r,f}$ | 0,3 | 0,17 | 0,25 | 0,083 | 0,083 | 0,17 | 0,17 |
| | | | | | | | | |
| Yang | Délka slova | 4 | 5 | 6 | 7 | 8 | | |
| | $d_{r,f}$ | 0,17 | 0,17 | 0,083 | 0,083 | 0,083 | | |

Tabulka 4-4 Výsledné R-F vzdálenosti sekvencí savců

| **Savci- kompletní mitochondriální DNA** | | | | | | |
|---|---|---|---|---|---|---|
| Četnost dimerů | $d_{r,f}$ | 0,9 | | | | |
| | | | | | | |
| Qi | Distance | 1 | 10 | 20 | 30 | 50 |
| | $d_{r,f}$ | 0,66 | 0,66 | 0,66 | 0,62 | 0,62 |
| | | | | | | |
| Yang | Délka slova | 4 | 5 | 6 | 7 | 8 |
| | $d_{r,f}$ | 0,59 | 0,483 | 0,45 | 0,41 | 0,38 |




Obrázek 4-4 Fylogenetický strom s užitím metody Yang (délka slova 8), sekvence savců

# ZÁVĚR

Nejprve jsem provedla analýzu sekvencí DNA u 13-ti primátů a 30-ti savců klasickým přístupem. Použila jsem k tomu funkci *multialign* pro vícenásobné zarovnání, funkci *seqpdist* k vytvoření distanční matice metodou Jukes- Kantor a výsledný krok analýzy, fylogenetický strom, jsem vykreslila pomocí funkce *seqlinkage* s použitím metody UPGMA. Poté jsem vytvořila jednoduché algoritmy klasifikace organismů založené pouze na porovnání četností nukleotidů, dinukleotidů i tripletů pomocí Euklidovské vzdálenosti a dvě složitější numerické metody Qi a Yang. Výsledné fylogenetické stromy jsem sestrojila opět pomocí funkce *seqlinkage* metodou UPGMA.

Když porovnáme účinnost porovnání samotných nukleotidů, vidíme, že se podařilo přibližně správně separovat pouze skupinu pěti organismů ( *Homo sapiens, Homo sap. Neanderthalensis*, *Pan paniscus, Gorilla Gorilla* a *Pan troglodytes*) a správně byl též vykreslen *Eulemur fulvus*, který stojí evolučně nejdále od všech ostatních. Zbylé organismy byly zařazeny značně nepřesně. Klasifikace pomocí dinukleotidů již poskytuje lepší výsledky. Podařilo se identifikovat dvojici *Homo sapiens, Homo sapiens neanderthalensis* i *Pan troglodytes, Pan paniscus* a dále i celou skupinu tří gibonů (*Hylobates lar, Nomascus leucogenys* a *Nomascus gabriellae*) i když s mírnou nepřesností. Podařilo se i správně určit vzdálenou příbuznost *Eulemura fulvus* ke všem ostatním. Předpoklad byl takový, že nejlepší výsledky z vytvořených jednoduchých algoritmů, co se týče účinnosti bude poskytovat porovnání tripletů. Toto se bohužel vůbec nepotvrdilo. Klasifikace tímto postupem se zdá být naopak nejvíce nepřesná. Zdánlivě správně identifikuje blízkou příbuznost *Homo sapiens* a *Homo sapiens neanderthalensis,* ale stejnou míru příbuznosti přiřazuje i k *Pan troglodytes. Eulemur fulvus* se opět nachází v nejvíce vzdálené větvi, ovšem i s paviánem (*Papio papio*) a orangutanem (*Pongo abelii*), který je navíc určen jako jeho velmi blízký druh. Z vytvořených metod se tedy nejvíce osvědčila klasifikace na základě dinukleotidových četností, proto jsem ji zvolila jako zástupce " nových" jednoduchých metod do vytvořeného programu.

Z testovaných složitějších numerických metod Qi a Yang se ukázala jako efektivnější druhá z uvedených, jelikož poskytla přesnější analýzu současně s kladením mnohem menších výpočetních nároků. Všechny metody vykazovaly lepší výsledky při analýze sekvencí primátů, jelikož byly jejich sekvence téměř stejně dlouhé a obsahovaly přibližně desetkrát méně nukleotidů. Díky těmto vlastnostem použitých sekvencí primátů se jako výpočetně vhodnější nástroj k jejich analýze jeví klasický přístup s využitím vícenásobného zarovnání. Jelikož je však mnohdy nutné klasifikovat i

skupiny zcela odlišných organismů- a tedy zpravidla i zcela odlišných sekvencí i jejich délek (jak je tomu i v použité skupině savců), vícenásobné zarovnání se stává téměř nerealizovatelným a tehdy jsou numerické metody jednoznačně efektivnější volbou. Snaha vytvářet stále nové a účinnější postupy tedy i nadále pokračuje, a pokud tak již neučinili, zcela určitě si v bioinformatické analýze najdou své pevné místo.

# CITOVANÁ LITERATURA


1] R. O. S. a. kol., 2012. [Online]. Available: http://www.zoologie.frasma.cz/fylogeneze/fylogeneze_C.html.

2] p. J. Zrzavý, Fylogenze živočišné říše, Scientia, 2013.

3] S. a. k. Rosypal, Fylogeneze, systém a biologie organismů, Praha: SPN , 1992.

4] Univerzita Karlova v Praze, Katedra experimentální biologie rostlin, [Online]. Available: http://kfrserver.natur.cuni.cz/.

5] P. D. I. T. Urban, „Mendelova univerzita v Brně, Ústav morfologie, fyziologie a genetiky zvířat," 2013. [Online]. Available: http://user.mendelu.cz/urban/vsg1/molekul/mol_genome1.html.

6] P. R. V. Martínek, „Portál Přírodovědecké fakulty Univerzity Karlovy na podporu výuky chemie," [Online]. Available: http://www.studiumchemie.cz/materialy.php.

7] P. M. e. M. M. Vácha, Ph.D., „Projekt lidského genomu," 2010.

8] Fakulta vojenského zdravotnictví Univerzity obrany, „Bioinformatika".

9] M. J. Pláteník, Ph.D., „Ústav lékařské biochemie 1. Lékařské fakulty UK, Sekvenování genomů," [Online].

10] „Deoxyribonucleic Acid (DNA)," National Human Genome Research Institute, 2012.

11] N. O. a. kolektiv, Obecná biologie pro lékařské fakulty, Jinočany: H&H, 2000.

12] „Aminokyseliny," 3. Lékařská fakulta Univerzity Karlovy.

13] P. G. Higgs a T. K. Attwood, *Bioinformatics and Molecular Evolution,* Blackwell Publishing.

14] P. M. A. Svoboda, CSc., *Translace a metabolismus proteinů.*

15] X. Qi, E. Fuller, Q. Wu a C.-Q. Zhang, *Numerical Charakterization of DNA Sequences Based on Dinucleotides,* The Scientific World Journal, 2012.

16] *Mutace a dědičné choroby,* Brno: LF MUNI, Přednášky z předmětu AMOL: Úvod do molekulární biologie, 2011.

R. P. Řehulka, Ph.D., *Bioinformatika.*
17]

F. Cvrčková, Úvod do praktické bioinformatiky, Praha: ACADEMIA, 2006.
18]

*Prezentace z předmětu Bioinformatika,* Brno: FEKT VUT, 2011.
19]

FEKT VUT, *Prezentace předmětu Bioinformatika.*
20]

Virtual Amrita Laboratories Universalizing Education, Alignment of
21] Sequences, 2013. [Online]. Available: http://amrita.vlab.co.in/?sub=3&brch=274&sim=1433&cnt=1.

B. L. Sliž, *Určování genetické odlišnosti biologických sekvencí DNA,* Brno:
22] FEKT VUT, 2013.

Přírodovědecká fakulta Univerzity Palackého, [Online]. Available:
23] http://www.dnabased.com/Bioinformatika/Analyza_nukleotidove_sekvence/index.html.

P. T. M. Lowe, „The Lowe Lab," [Online]. Available:
24] http://lowelab.ucsc.edu/.

T. Hasíková, Mechanismy udržování telomer u Řas, Brno: MASARYKOVA
25] UNIVERZITA V BRNĚ, Ústav experimentální biologie, 2011.

„Katedra ekologie a životního prostředí, Přírodovědecká fakulta UP;
26] Technologie klonování genů," [Online]. Available: http://www.ekologie.upol.cz/.

Dr a D. R. J. Edwards, University of Southampton, [Online]. Available:
27] www.southampton.ac.uk.

X. Yang a T. Wang, „A novel statistical measure for sequence comparison on
28] the basis of k-word counts," *Journal ofTheoreticalBiology,* 2013.

M. Randić, X. Guo a S. C. Basak, *On the Characterization of DNA Primary*
29] *Sequences by Triplet of Nucleic Acid Bases.*

Y. Zhang, *A Simple Method to Construct the Similarity Matrices of DNA*
30] *Sequences,* Weihai: Shandong University at Weihai.

M. Z. Jelínek, Molekulární typizace bakterií mléčného kvašení, Brno: VUT,
31] 2010.

W. H. Day, Optimal algorithms for comparing trees with labeled leaves,
32] Journal of Classification.

[Online]. Available: http://www.quretec.com/u/vilo/edu/2005-
33] 06/Text_Algorithms/index.cgi?f=L5_Edit.

„Wikipedia," [Online]. Available:

34] http://en.wikipedia.org/wiki/File:BLOSUM62.gif.

[Online]. Available:
35] http://wwwabi.snv.jussieu.fr/~erocha/webthese/bioinfoI.html.

„Wikipedia,“ [Online]. Available: http://cs.wikipedia.org/wiki/Soubor:Smith-
36] Waterman.jpg.

P. M. R. Průša, CSc, „Sekvenování nukleových kyselin,“ 2. Lékařská fakulta
37] Univerzity Karlovy.

[Online]. Available: bioweb.uwlax.edu.
38]

Jihočeská univerzita v Českých Budějovicích, „BRIDGE4INNOVATION,“
39] 2010. [Online].

M. Teplá, PřF UK v Praze, [Online]. Available:
40] http://www.studiumbiochemie.cz/translace.html.

# SEZNAM PŘÍLOH

# A ZDROJOVÉ KÓDY

## A.1 matice_distanci

```
%vypocet matice distanci pomoci normalizovane hodnoty
podobnosti
function [distance]=matice_distanci(sekvence)
zarovnane=multialign(sekvence);%multi zarovnani sekvenci
showalignment(zarovnane)%zobrazeni zarovnani
s=size(zarovnane);
r=s(1)%kolik obsahuje vstupni soubor sekvenci
distance=zeros(r,r);
for i=1:r
    for j=i+1:r

distance(j,i)=sum(zarovnane(i,1).Sequence~=zarovnane(j,1).S
equence)/length(zarovnane(1,1).Sequence);
        %vypocet normalizovane hodnoty podobnosti
    end
end
distance;%vysledna matice distanci
```

## A.2 baseMatice

```
function [distancebase] =baseMatice(sekvence)
s=size(sekvence);
s=s(1);
MatVek=zeros(s,4);%pomocna matice pro vektory cetnosti
for i=1:s
base=basecount(sekvence(i,1).Sequence);
cel=struct2cell(base);%za sebou hodnoty cetnosti nukleotidu
Vek=[ ];
for o=1:4
Vek=[Vek cel{o}]%za sebou hodnoty cetnosti nukleotidu
end
MatVek(i,1:4)=Vek;%matice vekt po sebou, radek = jedna
sekvence
end
distancebase=zeros(s,s);%prostor pro matici distanci
for i=1:s
    for j=i+1:s
        for k=1:4

distancebase(j,i)=sqrt(sum(((MatVek(i,k)/length(sekvence(i,
1).Sequence))-
(MatVek(j,k)/length(sekvence(j,1).Sequence)))^2));
        end

    end
end
```

## A.3 dimerMatice

```
function [distancedimer] =dimerMatice(sekvence)
s=size(sekvence);
s=s(1);
MatVek=zeros(s,16);%pomocna matice pro vektory cetnosti
for i=1:s
dimer=dimercount(sekvence(i,1).Sequence);
cel=struct2cell(dimer);%za sebou hodnoty cetnosti
nukleotidu
Vek=[ ];
for o=1:16
Vek=[Vek cel{o}]%za sebou hodnoty cetnosti nukleotidu
end
MatVek(i,1:16)=Vek;%matice vekt po sebou, radek = jedna
sekvence
end
distancedimer=zeros(s,s);%prostor pro matici distanci
for i=1:s
    for j=i+1:s
        for k=1:16

distancedimer(j,i)=sqrt(sum(((MatVek(i,k)/length(sekvence(i
,1).Sequence))-
(MatVek(j,k)/length(sekvence(j,1).Sequence)))^2));
        end

    end
end
```

## A.4 codonMatice

```
function [distancecodon] =codonMatice(sekvence)
s=size(sekvence);
s=s(1);
MatVek=zeros(s,64);%pomocna matice pro vektory cetnosti
for i=1:s
codon=codoncount(sekvence(i,1).Sequence);
cel=struct2cell(codon);%za sebou hodnoty cetnosti
nukleotidu
Vek=[ ];
for o=1:64
Vek=[Vek cel{o}]%za sebou hodnoty cetnosti nukleotidu
end
MatVek(i,1:64)=Vek;%matice vekt po sebou, radek = jedna
sekvence
end
distancecodon=zeros(s,s);%prostor pro matici distanci
for i=1:s
    for j=i+1:s
        for k=1:64
        distancecodon(j,i)=(sum(abs(MatVek(i,k)-
MatVek(j,k))))/((length(sekvence(i,1).Sequence)+length(sekv
ence(j,1).Sequence))/2);
        end

    end
end
```

## A.5 Qi

```
function Vzdal_city=Qi(sekvence,mez)
s=size(sekvence);
s=s(1);%pocet sekvenci
for k=1:mez%maximalni distance
    MatCet=[];
    for i=1:s %bere jednotlive sekvence
    sek=sekvence(i,1).Sequence;
    count=DimerCount(sek,mez);
    MatCet=[MatCet;count];
    end
    dist=pdist(MatCet,'cityblock');%distance
    b=size(dist);
    Vzdal(k,1:b(2))=dist;
end
  for o=1:b(2)
      Vzdal_city(o)=sum(Vzdal(:,o))/(mez);%prumerovani
  end
```

## A.6 Yang

```
   function Vzdal_eucl = Yang(sekvence,delka)
s=size(sekvence);
s=s(1);%pocet sekvenci
for k=2:delka %ruzne delky slov
kword = unique(nchoosek(repmat('ACGT', 1,delka), delka),
'rows');%vytvoreni vsech slov
por=[];
vekt=zeros(1,4^delka);
mat=[];
VysVekt=zeros(1,4^delka);
Sor_vekt=[];
for i=1:s %bere jednotlive sekvence
    sek=sekvence(i,1).Sequence;
    for j=1:4^delka
find=findstr(sek,kword(j,:));
siz=size(find);
siz=siz(2);%cetnost j-teho slova
vekt(j)=siz;%vektor cetnosti vsek slov 1 delky
    end
    Sort_vekt=sort(vekt);%serazeny vektor vzestupne
    for h=1:4^delka
    vys=findstr(vekt(h),Sort_vekt);
    VysVekt(h)=vys(1);
    end
    mat=[mat;VysVekt];%1 radek ukazatel poradi vyskytu
end               % slov 1 delky 1 sekvence
Mat=mat./4^delka;
euc=pdist(Mat,'euclidean');%distance
b=size(euc);
Vzdal(k,1:b(2))=euc;
end
  for l=1:b(2)
      Vzdal_eucl(l)=sum(Vzdal(:,l))/(delka-1);%prumerovani
  end
```